Commentary on Daniel Perttu's "A Quantitative Study of Chromaticism"

Art Samplaski[1]
Ithaca, NY

ABSTRACT: The methodology used in Daniel Perttu's article is analyzed for conformance to several criteria needed in quantitative studies. A number of problems are identified. Some of these appear to be deep structural issues given the nature of the question studied while others are caused by the methodology itself, by both the types of analyses carried out and the nature of the data source. Various suggestions to strengthen the study are made.

Submitted 2007 March 29; accepted 2007 March 29.

KEYWORDS: *chromaticism, melody, statistics*

DANIEL Perttu (2007) takes on one of the "received bits" of (folkloric) music wisdom: that at least since 1600, Western art music has been becoming increasingly chromatic. He takes a large sample of melodic data from Barlow & Morgenstern's (1948; henceforth, B+M) catalogue of themes and carries out several tests to check the proposition on both inter- and intra-composer levels. His use of statistical analysis is most welcome—this is exactly the type of question in historical musicology that is amenable to the use of quantitative methods, and it is delightful to see someone applying them. Unfortunately, an analysis of his methodology shows several problems, so that for the present we must render the Scottish verdict of "not proven." The case is not completely closed, however: a reworking of the methodology might tighten things up sufficiently that we could accept the assertion with confidence. (For the record, I hope that he succeeds, since my own musical intuition agrees with many others that the proposition is true.)

REQUIREMENTS FOR QUANTITATIVE ANALYSIS

A couple of paragraphs on the nature of statistical analysis are needed for readers not well-versed in the subject for them to follow fully much of this critique. Those who already know something of statistics can skip ahead.

To successfully apply inferential statistics to a question requires that three criteria be satisfied: 1) Was the question for which data were gathered well-defined? 2) Were there any problems with the data gathered that could bias any test results? and 3) Were appropriate tests applied? Criterion one is straightforward: if your hypothesis is not well-formulated (and by extension, your definition for data), there is no way to check that you are gathering relevant information. Criterion two is more complex in the details but still conceptually straightforward. If one went to take "an opinion poll of Canadians" but only questioned people in Toronto, then the results would almost certainly be unrepresentative of the overall population—a trivial example, but the same principle applies for far less obvious situations. Perttu's article is a case study of some of the subtleties involved.

Criterion three is the most technical and which makes many people run in horror at the mere word "statistics," but it is still simple in origin. Inferential statistics is always trying to answer a basic question: do two or more groups differ with respect to some characteristic? That question is very complicated to answer because data can vary so dramatically in make-up, and each individual statistical test must make particular assumptions about the data. If the data do not meet those assumptions, the results are almost certain to be invalid. Hence, many different tests are needed to cover all the cases.[2] Let us examine the extent to which Perttu's report meets these criteria.

WELL-FORMEDNESS OF THE HYPOTHESIS' DEFINITION

Perttu acknowledges up front the potential danger involved in his operational definition of "chromaticism," and he is to be commended for it. The main pitfall, though, is mentioned only in passing: the difference between chromaticism in the melodic vs. harmonic dimensions. If one only looks at melodies and not their harmonizations, an important and possibly key component of era-specific chromaticism will be missed. For example, Martino's (1984) comparative edition of Bach chorale settings shows in top-to-bottom format

every instance where Bach did multiple harmonizations of a chorale tune. The tunes themselves are quite diatonic, but the settings differ significantly in the amount of chromaticism in other voices—something missed entirely by Perttu's operational definition. To be sure, it would be difficult to incorporate the harmonic dimension of music into a good operational definition for a study of this sort, and I do not blame Perttu for side-stepping the problem. There is a dissertation topic waiting here for anyone courageous enough to try.

By comparison, the failure to distinguish different degrees of chromaticism (passing and neighbor vs. structural tones) seems much less worrisome. If melodies of an era contain 65% chromatic tones, the end effect is likely to be the same whether those tones were structural or passing/neighbor ones. The one possible disagreement about whether notes are "structurally chromatic" is how to categorize raised scale degrees 6 and 7 for ascending melodic minor—are they really outside the key, even though they require accidentals? Many careers have been spent on this problem (see Jorgenson [1957]), so it seems best to leave the issue alone.[3]

An entirely separate issue is the meaning of the words "diatonic" and "tonal." Any number of modally-organized or neotonal melodies might fall afoul of this definition of chromaticism. Since a discussion of this issue would go far beyond the scope of any single article or critique, let us acknowledge but side-step it by saying that Perttu is investigating the level of melodic chromaticism over time vis-à-vis "classical tonality."

PROBABLE BIASES IN THE DATA

Perttu says that he used B+M as his data source because it was a readily-available catalogue of a large (ca. 10,000 items) body of melodies. He states one problem with the data—uncertainties in the dating of pieces—and details his method for handling it. There are, however, other issues that he does not mention, whether or not he attempted to deal with them. These include: 1) omission of vocal music; 2) overweighting in favor of specific composers; 3) overweighting of specific time periods; 4) potential systemic differences in melody length as a function of era; and 5) nonindependence of the data.

First, B+M's sample of themes must perforce be highly biased since, in their own words (p. xi), "[The themes] have been chosen primarily from recorded, instrumental pieces. No vocal works, excepting those which in instrumental arrangement have become better known than their originals, have been included." Major figures such as Monteverdi, Schütz, and Puccini are thus missing entirely; composers like Verdi, Rossini, Donizetti, and Bellini are confined to a few overtures each; and even well-represented composers are lacking major works or large portions of their output—no Bach cantatas or oratorios, Brahms or Mozart Requiems, Schubert *Lieder*, ad nauseam. Meanwhile, B+M go to the opposite extreme for a number of included works: Brahms' Symphony no. 1 has 25 entries, for example. Their own definition of "theme" thus does not seem particularly well-formulated.

As an immediate corollary, B+M's sampling criterion means the collection must consist primarily of "melodies of the great masters." While they net a fair number of "second-tier" individuals and occasionally pick up pieces by truly minor composers—Julius Fucîk is not in any way a household name, for example, but his "March of the Gladiators" is instantly recognizable by anyone who has ever attended or seen a film of a circus—the anthology is highly skewed towards the set of canonical composers. This is very filtered as compared to the set of all composers active during the period 1600-1948.

Perttu could have slightly mitigated these specific issues by restricting his operational definition to "melodic chromaticism in instrumental music by canonical composers." However, for many readers this could undermine any conclusions he reached. Even if the hypothesis as roughly formed in people's minds is implicitly defined in terms of the canon, the lack of vocal music might be considered too major an omission of data. It is not at all clear how the level of melodic chromaticism varies over time in vocal music versus instrumental music. A movement like the Kyrie fugue of Bach's B-minor Mass is highly chromatic, but that might be an anomaly; if musicians asked colleagues to name a dozen vocal themes by major composers, it is fair to say that most would likely expect to get replies like Handel's "The Trumpet Shall Sound" or Mozart's "Si vuol ballare, Signor Contino." The possible difference in degree between vocal and instrumental chromaticism is in fact an entirely separate question and requires its own quantitative study. Meanwhile, let us move on to the other issues.

The next problem involves sample sizes within B+M, and has two components. First, unequal lengths of time are covered by both the sample and the standard historical categories. B+M's book spans the entire 17th, 18th, and 19th centuries, but 20th century coverage stops ca. 1947. Again, the nature of the data source forces this: B+M could hardly be expected to include melodies composed decades after their book went to print. The second component is the number of melodies included per historical category:

Perttu uses 558 themes over the 150 years of the Baroque period vs. 1110 for the 70-year long Classical period vs. 3347 over the 90 years of the Romantic period vs. 794 in the 30-odd years of the post-1911 Modern period. Clearly, the sample is heavily overweighted towards 19th century melodies.

There is a simple workaround to this two-part problem: ignore the standard historical periods, and instead choose equal numbers of themes at random within time periods of equal length, ending with B+M's publication date. Perttu partially does this for the data in his Figure 2, which looks at 20-year time slices. He unfortunately still does not use equal numbers of melodies for those periods (e.g., 1207 themes for the period 1880-1899 vs. 67 for the period 1700-1719). Had he randomly sampled 30 or 40 themes per bi-decade rather than use every melody identified for them, his results would have been stronger.[4] Of course, this workaround is in turn affected by the melody dating problem.

B+M do not list dates except for works under copyright at the time, so Perttu had to obtain such information from other sources. He says that he used the cautionary approach of leaving out any melody with posthumous publication. This has a substantial risk of underrepresenting canonical composers like Schubert. Even without this, there are cases where publication date is misleading: for a number of Brahms' compositions, for example, the opus number (and hence the publication date) has little to do with the order (and year) of composition.

Next, there might be some type of subtle but systematic bias due to lengths of melodies. Perttu states that the average melody length was 19.4 notes, but he did not conduct any check against the possibility that melodies from different eras are generally of different lengths (Perttu, personal communication). If Romantic melodies are systematically longer than Baroque melodies, there are already more opportunities to find chromatic notes in the former, whether or not there is a structural increase in the level of chromaticism between the two eras. If there were a systematic difference in melody length by era, that would require additional safeguards in the sampling of data. Once again, this is its own separate question that requires a prefatory investigation.

The final and most difficult problem: the data are not independent. No composer in beginning a new work is ever overcome by total amnesia about their prior oeuvre so as to start tabula rasa. Such influence by composers' own prior works, let alone knowledge of the compositions of their contemporaries and predecessors, is a major problem for most statistical tests, since they have as a primary assumption that all observations are independent and uninfluenced by each other. This issue affects all of Perttu's tests, and is detailed below.

APPROPRIATENESS OF TESTS AND IMPLICATIONS FOR RESULTS

Perttu's initial test is a correlation—does the percentage of chromatic notes increase with date of composition? There is nothing wrong in principle with this; it is the simplest way to check for any gross effect. The problem here is an issue called "statistical significance" vs. "practical significance."

Perttu reports a correlation value r of –.155 (diatonicism decreases as composition date increases), of which he states (2007, p. 49), "While the magnitude of this correlation would seem to be modest, it proves to be statistically significant ($p<0.0000000000000001$)." The p-value is the probability that one might have obtained this result by chance, so prima facie this appears to settle the question: a 1 in 10 quadrillion chance of a false positive seems like an extremely safe bet, to say the least. It fails, though, to consider that for any suitably large sample of data, it becomes very easy to find "statistically significant results" that do not explain any useful amount of the variability of the data, i.e., the result has no practical significance.[5]

A personal cautionary example is illustrative. In the pilot work for my dissertation, I had to determine whether or not the data from different participants should be included in the analysis. For the test being used, the usual criterion was, include the participant's data if it correlates with the averaged data of the entire population sample at $p<.05$ (i.e., less than a 5% chance of a false positive). As a safety, I had included some duplicate trials at random as a self-consistency check for each participant. One person had a self-consistency correlation of $r=.007$—her responses were essentially noise. Yet, taken over the entire 240 real trials she still correlated with the sample average at $p<.001$, solely because of the number of trials involved.

Here we are dealing with a sample of 5809 melodies. Regardless of any problems with the data, we are almost guaranteed to obtain a statistically significant result just from sheer sample size. Squaring the r-value tells what percentage of the overall variability of the data is being accounted for by the correlation. In this case, $r^2=.024$, so that less than 2.5% of the variation in the data (the overall amount of chromaticism across all the melodies) is explained by an increase in the level of melodic chromaticism over time. By itself, then, the correlation test does not provide much practical value, and we need to discover

other factors that might account for the other 97.6% of variation.[6]

Perttu next analyzes chromaticism across conventionally-defined style periods and 20-year time slices (Figures 1 and 2). To compare the conventional style periods he used something called a t-test. This is a very basic test to determine if two population samples have different means (averages) for some characteristic. It is only appropriate for comparing two groups, however—Perttu is comparing four groups in Figure 1 and 13 groups in Figure 2. Using t-tests in such a situation means that the risk of reporting a wrong result goes up dramatically. To compare four groups, for example, requires six pairwise comparisons. Suppose one is willing to accept a 5% chance of a false positive result—the technical term is "sets the alpha-level at .05"—on one's t-tests.[7] This means that each t-test has a 95% chance of giving a correct result; but the overall probability of obtaining a correct result is the product of the probabilities of the individual tests. The effective alpha-level (the total probability of a false positive) goes up very fast as the number of groups increases. For four groups it is approximately 3 in 10, and by six groups one has about a 3 in 4 chance of a false positive. Perttu does not discuss any tests done for Figure 2; with 13 groups over which to do pairwise comparisons (70 are needed) he would have over a 97% chance of obtaining a false positive if he used t-tests.

The appropriate test here is an analysis of variance, or ANOVA for short, a generalization of the t-test that allows one to compare simultaneously more than two groups while avoiding the above problem. There are many types of ANOVA, depending whether various assumptions about the data hold true. The simplest type, for example, is very badly affected if the data are not independent; and Perttu has indicated (personal communication) that this was a reason he used t-tests. However, there are other kinds of ANOVA designed for this situation, and one of them needed to be used.

The situation is complicated by the two kinds of dependency in the data: intra-composer (composers' knowledge of their own work) and inter-composer (knowledge of predecessors' and contemporaries' work). There is likely no way to eliminate completely any effects due to the latter[8] and its implications for analysis are difficult to understand fully; but the former can be mitigated somewhat. All melodies from each composer during each 10- or 20-year period can be averaged together as a lump sum, obtaining a "group chromaticism score" for the composer. The unit of observation would then be "chromaticism of individual composers per (bi)decade" rather than "chromaticism of individual melodies." There would still be the problem of correlated observations due to the length of composers' careers—many composers would be present in five or more decades—but it would be a better starting point.

Perttu's second study yields null results, i.e., no significant increase in chromaticism over the course of five composers' careers. Those results might be surprising, but the reported *p*-values are all well beyond being borderline.[9] While he tries to suggest that there might some weak effect for Brahms, this simply cannot be taken seriously. Beyond the high *p*-value of .165, the r^2-value—the amount of variation in the data being accounted for—is .0025, so only one-quarter of 1% is covered. Perttu discusses some possible causes for these results, focusing on Beethoven. In particular, he considers the possibility that chromaticism in Beethoven occurs not in the primary melodic line but in subsidiary parts. This goes back to his concerns over the proper operational definition of chromaticism, and definitely deserves investigation. Aside from that, however, the problems with B+M as a data source may well be so great as to preclude feeling secure about any result one way or another. The study needs to be redone from scratch.

The final study also finds a null result, but it deserves a comment. A number of readers might feel disturbed by the "thrown-dart" method used to collect the data. They might think that this random sampling of isolated notes fails to capture any meaningful use of chromaticism by Mozart. However, this data collection method is actually the most robust, because it is the least susceptible to any subtle systematic problems. For example, the notes within a measure are highly correlated with each other. If random measures were sampled rather than random notes, there would certainly be systematic biases. While nonindependence of observations remains a background issue—Mozart was influenced by his prior works—the dataset as developed is of a type known as a "totally-within" design, and there are tests suited for analyzing such data. It is the most bias-free method for data collection used by Perttu, and eliminates most of the problems caused by using B+M as data source. As such, it could be used in a replication of his second study, particularly for Beethoven.

A general note about the bar graphs in the article: Perttu fell into the trap of trying to condense his figures, which creates misleading diagrams. For example, the columns "Baroque" and "Classical" appear to be twice as high as the "Modern" column in his Figure 1, whereas if one examines the vertical scale, the difference is actually 94% vs. 86%, a very different ratio. Similar problems exist for Figures 2 and 3. This type of diagramming error is one of the standard disinformation techniques detailed in Huff's (1954) classic *How to Lie with Statistics*, a book that should be required reading for every consumer. Let me be absolutely clear that I am *not* suggesting Perttu deliberately tried to be misleading—this was an innocent

error. It shows, though, another example of how easy it is to have problems in a quantitative study.

CONCLUSIONS

Perttu's analyses need to be rerun using appropriate tests applied to much cleaner data before we should consider accepting his results, however much those results might agree with our own intuitions. I repeat my sentiment from the introduction: I hope that this is done some day because I, like many other musicians, think that Western art music *did* become more chromatic over the common-practice period, on the basis of my own experience with that repertoire. Anecdotal evidence, though, is not sufficient to believe a broad assertion. There are far too many instances where "everybody knows X is true" only to be shown that it is not so once some actual figures have been collected. A quantitative hypothesis must be tested by empirical methods. If a replication of Perttu's study, done with stricter methodology, finds that chromaticism did not significantly increase over that time-span, I for one would be quite surprised—but I would accept the result.

A number of readers might conclude that this type of analysis is too complicated to be worth using. Worse, they might think that this demonstrates why "we should keep scientific methods out of the arts." The latter is an "ostrich with its head in the sand" attitude and deserves contempt. It is certainly clear that statistical studies are not appropriate for many questions in historical musicology: analysis of composers' diaries and correspondence to try and gain some insights into their psychological states is but one example. Other questions do warrant the use of such tools, though. For example, attempts to assign authorship of works or to verify the authorship of doubtful works on the basis of analysis of the music itself are quantitative processes, and have long been carried out in an implicit (and as a result, possibly crude) manner by historical musicologists. To read that "this type of melodic pattern is highly (un)representative of Haydn," is to read a statistical assertion about relative frequencies. It is, however, an anecdotal assertion, rather than being backed up by tests that give numerical margins of error.

Several humanities areas have begun using quantitative methods where appropriate for their field. Journals of government/political science, for example, now routinely carry articles that make detailed use of statistics, and the results reported therein can have significant policy implications. A case in point is Wand, et al's (2001) thorough and rigorous analysis of the voting problems in the highly-disputed 2000 U.S. presidential election, proving that the problems with the infamous "butterfly ballots" in Palm Beach County, Florida, were the cause of the court-declared loss of Gore to Bush. More immediately relevant for music analysts, if less far-reaching, is Huron's (2001) study of motives in Brahms's Op. 51, no. 1, which forcefully argues against an earlier analysis by Forte.[10] An earlier quantitative application in music was McHose's (1947) manual on part-writing, where he and his assistants compiled detailed frequency counts of chord-progression types used by Bach in his chorale settings, and used those as the basis for guidelines to students. While the book was greeted with some distaste at publication, it gave a much firmer foundation to the pedagogy of part-writing.

Again, readers might feel some dismay at the prospect of having to learn an entirely different (and probably somewhat alien) methodology in addition to the other required tools of historical or systematic musicology in order to do their research. They might feel particular unease considering the number and types of problems found in Perttu's data source. How, they might wonder, can we possibly have sufficiently clean data for any empirical study to be valid?

There are several reassurances. First, not every music analyst would need to learn statistics, just as someone studying works of the second Viennese school does not need to learn codicology or an Ars nova specialist does not need to know about atonal similarity functions. Second, the problems noted with B+M as a data source are somewhat of a "perfect storm" given the particular broad question Perttu was investigating. Someone doing a McHose-type stylistic analysis of a composer or a Huron-style motivic analysis of a few pieces would face far fewer worries about their datasets. Third, while any music analyst contemplating the use of quantitative methods in their research needs to learn enough statistics to forestall problems arising from any "a little knowledge is a dangerous thing" situation, that does not mean that they must spend several years learning the subject. In particular, they should consider the wonderful opportunities for interdisciplinary collaboration with researchers who already have considerable statistical expertise. Such collaborations might well generate innovative approaches for the types of questions being investigated, due to individuals with fresh perspectives coming into the process. Finally, music analysts should not shy away from the potential for controversies over datasets. Arguments over data, results, and their interpretations have always been part of scholarly debate, whether in the humanities or the sciences. Such byplay will continue independent of the use of new methodologies.

ACKNOWLEDGMENT

I am extremely grateful to the Editor of *Empirical Musicology Review* for making it possible to communicate with Mr. Perttu while preparing this commentary, in order to clear up several points.

NOTES

[1] The author may be reached at "agsvtp" [at-sign] "hotmail" [dot] "com" (written this way to foil spam-trawling robots).

[2] An equivalent situation in mathematics is, try to prove the proposition "Any closed curve in a plane partitions the plane into two disjunct regions." This is visually obvious, but the question is so general that it is extraordinarily difficult to actually prove it.

[3] The possible confound of local modulations is a non-issue. Even if it were not rendered moot by B+M's listing of themes in the local key, it would "merely" be a problem in data entry to transcribe all melodies into their respective local key.

[4] Presumably Perttu would also want to leave out any post-1939 melody, since B+M stops ca. 1947—his final "1940-1959" bin would be biased through incompleteness.

[5] The value as reported amounts to overkill: no one reports a p-value with 15 leading zeros. If one obtains a $p<.001$ (or at most $p<.0001$), that is sufficient for publication; no finer level of detail is needed, except perhaps for extremely restricted situations such as clinical drug trials. As reported, the number could possibly make readers feel like they are being subjected to a high-pressure sales pitch.

[6] One prime candidate for revisitation: what happens if we reclassify raised scale degrees 6 and 7 in ascending melodic minor as diatonic?

[7] This is the alpha-level used in many studies. Researchers can be more conservative and decide to use a tighter criterion, e.g., an alpha level of .01, to reduce the chance of obtaining a false positive. (One cannot, however, start by setting alpha to .01, just miss a significant result (say, $p=.02$) and then post-hoc decide to settle for an alpha of .05 instead. One must decide a priori what alpha level one needs and stick with it.)

[8] The effect of inter-composer influence is even more complex than one might think at first, given that a composer like J. S. Bach was neglected for several decades save by selected individuals. A number of Mozart's compositions after he was introduced to and studied Bach's works show a striking influence in terms of contrapuntal sophistication; had he lived longer, perhaps he would have been the one to lead the "Bach revival" instead of Mendelssohn. Meanwhile, the evidence for any influence by Bach on, say, Chopin or Schubert is scantier.

[9] There is always a danger in quantitative studies of committing a "Type II error"—reporting a false negative. However, researchers in general are (rightfully) more concerned about reporting false positives (Type I errors). There are no warning signs in Perttu's methodology that suggest a check against Type II error is needed.

[10] The article is in two parts. The first is a general and non-mathematical discussion of the question, "What constitutes an important feature of a musical work?" and is one of the most lucid analyses of a basic problem in music theory that I have ever read. The second half is the detailed statistical analysis applied to the Brahms piece.

REFERENCES

Barlow, H., & Morgenstern, S. (1948). *A Dictionary of Musical Themes*. New York: Crown. Repr. ca. 1975 with updates to composers' dates.

Huff, D. (1954). *How to Lie with Statistics*. New York: Norton. Repr., 1993.

Huron, D. (2001). What is a musical feature? Forte's analysis of Brahms's Opus 51, no. 1, revisited. *Music Theory Online*, Vol. 7, No. 4.
http://www.societymusictheory.org/mto/issues/mto.01.7.4/mto.01.7.4.huron.html

Jorgenson, D. (1957). *A History of Theories of the Minor Triad*. Ph.D. dissertation. Indiana University.

Martino, D. (1984). *178 Chorale Harmonizations by Johann Sebastian Bach*. Newton, MA: Dantalian.

McHose, A. I. (1947). *The Contrapuntal Harmonic Technique of the 18th Century*. New York: F. S. Crofts.

Perttu, D. (2007). A quantitative study of chromaticism: Changes observed in historical eras and individual composers. *Empirical Musicology Review,* Vol. 2 No. 2, pp. 47-54.

Wand, J., Shotts, K., Sekhon, J., et al. (2001). The butterfly did it: The aberrant vote for Buchanan in Palm Beach County, Florida. *American Political Science Review*, Vol. 95 (December), pp. 793–810.